Student's Performance Analysisin Examination Using Machine Learning Technique

Submitted in partial fulfillment of the

Requirement for the award of a degree

Of

Doctor of Philosophy

In the

Department of

COMPUTER APPLICATION & IT

YBN UNIVERSITY

RAJAULATU, NAMKUM, RANCHI

JHARKHAND 834010 (INDIA)

January, Year 2021

Research Supervisor:

Dr. Manisha Kumari Deep
Designation: Professor
Department: Computer Application & IT

Research Scholar:

Name: Sana Jawed
Enrolment No:YBNU2000378
Department: Computer Application & IT

Abstract

Student’s Performance Analysis in Examination using machine learning system is developed to analyse and predict the student performance based on various criteria like student gender, race/ethnicity, parental level of education, their nourishment, test preparation course and their IQ level analysis by math , reading and writing scores. The proposed framework analyses the students demographic data, study related and psychological characteristics to extract all possible knowledge from students, teachers and parents. Seeking the highest possible accuracy in academic performance prediction using a set of powerful machine learning techniques called supervised learning using the algorithm of linear regression. This project is focused on improving data performance prediction based on their personal and academic performance characteristics.

Table of Contents	Page No
Abstract	2
Table of Contents	3
1. Introduction	4-5
2. Literature Review	5-14
3. Research Gap	14-15
4. Objectives	15-16
5. Scope of the work	16-17
6. Hypothesis	17-18
7. Research Methodology	18-24
8. Conclusion	24
9. References	25-27

1 Introduction

In the real world, we are surrounded by humans who can learn everything from their experiences with their learning capability, and we have computers or machines which work on our instructions. But can a machine also learn from experiences or past data like a human does? So here comes the role of machine learning.

Machine Learning is said as a subset of Artificial Intelligence that is mainly concerned with the development of algorithms which allow a computer to learn from the data and past experiences on their own. Student Performance Analysis System(SPAS) is an application which refers to other analysis systems like the car quality analysis, Share Market analysis and in broad perspectiveData Analysis. Student Performance Analysis in Exam using Machine Learning is developed to help teachers i.e. Faculty and Students enrolled and is useful for maintenance of records as well. Students have a significant impact on a country's social and economic growth and, a student's success is directly correlated with his/her academic performance. The advent of machine learning technology has led to vast amounts of educational data stored in different formats, requiring a proper method of extracting knowledge from large repositories for better decision-making.

Modern learning institutions operate in a highly competitive and complex environment. Thus, analysing performance, providing high-quality education, formulating strategies for evaluating the student'sperformance and identifying future needs are some challenges faced by most universities today.

The Machine Learning discipline deals with exploring, researching, and implementing Artificial Intelligence Techniques. As educational Artificial Intelligence is used to improve educational policymaking using data-based models, it has emerged as a vital research area to reveal presentable and applicable knowledge from large educational data repositories. Learning

analytics is a closely related field that deals with developing methods that lead to effective educational data sets to support the learning process.

2 Literature Review

This review is used to analyse student's performance in exam using machine learning based on past research in this field. It also identifies strengths in the existing literature and highlights the unique contribution that the study makes to learning analytics.

Several studies have been conducted to investigate the factors affecting student's performance using learning analytics and machine learning techniques. Academic determinants such as student qualification, parental education, their nourishment and their IQ level analysis by reading, writing and learning scores have affected student performance.

Kasi et. al. concluded that it is important to examine the factors that impact students' academic performance as it can improve teaching and learning processes[1]. Numerous academic and non-academic facets affect student performance, such as study time, frequency of absences, recreational activities, and interpersonal relationships. Educational data mining (EDM) and learning analytics (LA) are two closely related fields that reveal useful information from educational databases to generate actionable insights.

Durgaet. al. stated that student's scores across few important measures that tries to determine their academic excellence such as subjects conceptual knowledge based on marks scored in that exam, time management skills based on number of questions attempted is obtained priory, have been used to create homogeneous groups of students who opted for same subjects and share perceptions of their non-cognitive abilities [2]. Result of statistical graphical data helps students and teachers to track the performance and direction in

which the students perception and knowledge regarding the subject is inculcated. The results indicate the degree to which specific profiles predict outcome of the conceptually relevant data and widen the scope of their information to discover knowledge from data. Also it facilitates smooth and efficient handling of data that includes storage and modification of data in files which can be easily converted to other formats as desired by the user.

Reddyet. al. found out that automation to do mundane tasks and the expectations for students already equipped with good programming skills is on the rise [3]. In parallel, there have been a rising number of students who find it difficult to attain the skills necessary in order to get the dream IT job they desire. So they did the work to bridge the gap between the employer and the future employee of the company by the use of Student Performance Analysis System(SPAS) at college level. SPAS is an online web application system which enables students to know prior hand if their level of skills for the placement is enough to get placed or not, given the necessary inputs. SPAS have an intelligent learning algorithm which utilizes a rich database, analyses the records of previous student's traits and develops a model for further prediction. The performance evaluation of students by SPAS is by the cumulative predictor algorithm involving generation of several random forest trees on the available data. SPAS learn and create its model reaching higher accuracy with increasing data availability.

In one of the study Karthikeyanet. al. states that academic output of the student is normally stored in various formats in the student administration system (files, documents, records, photographs and other formats) [4]. These data can be collected for valuable knowledge from the students. However, it is difficult to analyse the increasing amount of data of students through conventional statistical techniques and database management tools. For universities to gather valuable information, a tool is therefore required. This

helpful knowledge can be used to predict the success of students. The analysis learning results is a framework that aims for success in the areas of student interest at various levels and dimensions. Their work proposes a complete structure as a rule-based recommendation method not only for analysing and forecasting the success of students but also for presenting their reasons. Their system analyses demographic details for pupils, studies and psychological features so that students, teachers and parents can collect all possible knowledge. To seek maximum accuracy in academic predictions across a range of powerful techniques of data mining the system successfully recognizes the limitations of the student and makes adequate recommendations.

Jirageet. al. used the data clustering technique named k-means clustering which is applied to analyse student's learning behavior[5]. The student's evaluation factor like class quizzes, mid and final exam assignment are studied. It is recommended that all these correlated information should be conveyed to the class adviser before the conduction of final exam. That study will help the teachers to reduce the drop out ratio to a significant level and improve the performance of student.

In one of the study Prajapatiet. aldid the survey of existing tools and techniques that have been designed [6]. Their work concluded that the main concern of the teachers regarding theperformance issue of the students is covered byintroducing the analysis tool R. Thisanalysis tool R uses machine learning approaches likeNaive Bayes, K-nearest neighbor (KNN) first for building amodel. Their work suggests thatone should use the cognitive modeling for designingthe knowledge-based so that they can produce betterresults.

Anushaet. al. mainly focused in their work on quality and value of the student's advancement in his profession as a challenge or an inquiry [7].These organizations center on quality in training consistently countless alumni from schools and colleges, as for the information gathered from the criticism of students, order an information mining strategy is connected to it. It is a stage to examine the elements influencing the performance of students so as to assess the present understudy execution and make proficient strides in the expectation of the no doubt happening connections between the different parts of learning and to upgrade the nature of instruction in future and help the educational organizers to design in like manner.

In one of the studies Kumar et. al. have used the combination of K-Means clustering algorithm with Artificial Neural Network(ANN) and Support Vector Machine(SVM) classification algorithm to evaluate the student performance in order to reduce the human effort [8]. The evaluation is done on the basis of Mean Square Error and Effort Estimation. The results of their implementation shows that the performance of Artificial Neural Network in comparison to Support Vector Machine is better. The Mean Square Error is 5-20% better whereas the Effort Estimation is around 15-27% better. The future possibilities include changing the total number of neurons or varying the satisfying parameters.

In one of the study Naveen et. al. have explored the machine learning algorithm on e-learning data has been proposed as a useful method to compare student’s performance as well as supporting certain prediction [9]. The performance evaluation of machine learning algorithm is also compared to determine the effectiveness for the task. This prediction can benefit teacher and course planner learns how to increase student’s results in the university exam. They have used cycle test mark, model mark as data sets. While for prediction techniques, the classification method is frequently used in educational data mining area. Under the classification techniques, Neural

Network and Decision Tree are the two methods used for predicting student's performance.Their future scope is that the system can be implemented as a web based application and android application so that the many of the institution may get benefits by using this system and many more features can be added which are related to school.

In one of the findings Pallathadkaet. al. anticipated that student's talents and interests might be linked to their performance [10]. A teacher's success is frequently measured by the performance of his students. Their work deal with dataanalysis from students and teachers. The analysis might be used tocategorize or forecast machine learning algorithms such as NaveBayes, ID3, C4.5, and Support Vector Machines (SVMs) which are a popular supervised learning algorithm used for classification and regression tasks. These algorithms are being researched in the experimentalinvestigation. A data set of UCI machinery student performanceis employed. Algorithms are evaluated based on characteristicssuch as accuracy and error rate. SVM is the most accurate techniquefor classifying a data set analysing student's performance.

In one of the study Acharya et. al. analysed the performance of the students and their attributes. The data set are derived bysurveying a group of students of colleges it is to be noted that the developed research methodology is general in nature [11]. It could be applied to full time as well as distance educationcourses including the courses where web based learning is done.The training set contains 309 instances whereas the testing setcontains 104 instances. After detailed analysis it was found that the most convenient algorithm to generate the setof production rules C4.5 was used to generatethe decision tree. F-Measure and Kappa Statistic were used todetermine the efficiency of the prediction algorithm. AverageF-Measure value for the training dataset was found to be 0.79whereas for the testing dataset was found to be 0.66. The latervalue seems to be slightly low perhaps due to the fact that

thenumber of testing instances is only 104. Similar results areshown by Kappa Statistic also.Priyaet. al. concluded that use of machine learning methods in conjunctionwith supervised learning algorithms better understand the efficiency of thealgorithm with respect to student records, where they analyse student performance andclassified it into three categories: high, medium, and poor, with a 79% precision [12]. They had some technological solutions in the future to increase the quality of studentsuccess. The user interface model was planned developed to automatically include studentrecords and to send staff warning messages about students who are doing poorly. They used a neural network to construct the prediction and anticipate improved performance.Non-academic qualities should be combined with academic attributes.In one study Dhilipanet. al. focuses on the student academic growth analysis using machine learning techniques [13]. Foranalysis Binomial logical regression, Decision tree, Entropy and KNN classifier were used. That processcan help the instructor to decide easily about performance of the students and schedule better methodfor improving their academics. In future additional features were added to their dataset to acquire betteraccuracy.

Alshabandaret. al. depicts that two sets of exterminates have been carried out in their study using regression and classification analysis [14]. The results of predicting student's assessments grades model shows that the student's performance in a particular assignment relies on student's mark in the previous assignment within single courses. The researchers conclude that student's prior grade point average (GPA) with a low mark is considered as a significant factor of withdrawal from the next course in the traditional classroom setting, The date of student deregistration from the course is a valuable predictor that is significantly correlated with student performance. With the regression analysis, the data does not provide the last date of

student’s activity prior to undertaken assessments. The finding results have been recommended to take into account the temporal features on predicting of subsequent assessments grades. Future research direction involves the use of temporalfeatures for predicting student’s assessments grades model. With temporal feature time series analysis will be untaken might be more advanced machine learning will be utilized.

In the study of Kotaiahet. al. states that feature engineering, the method of transforming dataset for machine learning, is including an essential factor in propagating reliable prediction outcomes [15]. Two diverse datasets were examined with three various machine learning techniques, and their outcomes were analysed applying for evaluation criteria. Techniques applied were Naïve Bayes classification, decision trees, and linear regression. For the computation of machine learning techniques, feature engineering was employed to the modified and raw version of the dataset individually. The central process of feature selection was feature engineering. In the state of regression and classification trees, supplementary feature engineering was prepared in the manner of custom feature production. Feature engineering was performed both with automatic interpretation and manual functionality of the dataset. In summation, features tuning was made with an error and trial strategy. The outcome of both datasets shows differences and relationships with their employment in fundamental education. In the first dataset, the relationship is that recall value was frequently greater than the precision value. The variation was in the efficiency value. This can be connected to the various individualistic attributes. In the original analysis, the dependent attribute was not changed into binary valued attributes. Concluding the dependent attribute might have addressed the predictions carefully in the research. The second dataset, the primary analysis adopted an additional attribute that symbolizes the prior exam ranks and delivered better

performance than from their studies. Though, once those attributes are neglected, accuracy rates were comparable to that paper.

In one of the study Kotaiahet. al. depicted thatfor predicting student's performance meta-analysis on predicting student's performance has motivated them to carry out further research to be applied in the proper environment [16].It will help the educational system to monitor the students' performance in a systematic way.

Memonet. al. explored and examined a massive VLE(Virtual Learning Environment)basedstudent behavioral dataset in terms of magnitude [17].Through literary studies of student behavioral studies, amaster feature vector is created possessing influentialproperties controlling the student's behavior. Twodifferent analysis procedures further examine thefeature vector, i.e., expert opinion is obtained foridentifying procrastination tendencies. In parallel a similar feature vector wasanalysed for automatic patterngeneration using K-means clustering. For an optimalnumber of cluster identification, the elbow method issued. A student behavioral dataset focusing onprocrastination behavior is proposed to representenough of the problem domain in terms of the datasetsize. The existing VLE dataset is enhanced by anadditional label to each record that identifiesprocrastination in students interacting with VLEs. Theprocrastination dataset is further fed to the classifierfollowing a detailed analysis protocol based on twoalgorithms. These classifiers are fine-tuned to getoptimal performance accuracy. The best performerclassifier hits the accuracy of 83.3%. In conclusion, thestudy was completed in the following aspects: 1)Theproposal of procrastination related student behavioral dataset in terms of considerable size.2) Detailed stepsin the form of algorithms for data formation and classification. 3) Detailed visualizations of clusteringand classification to identify procrastinationtendencies.

Guptaet. al. created an evaluation model was implemented using various machine learning techniques [18]. A student dataset of 500 records is taken for training and testing of KNN, SVM, decision trees, Naïve Bayes and KNN supervised learning methods. The work mainly focused on four primarily important features such as attendance, quiz, assignment and terms marks for evaluation of student's performance. These machine learning techniques classified the result data into three classes, namely excellent, good and improve. They also have compared these supervised learning techniques in terms of its accuracy rate and found SVM and KNN better as compared to others. As a further progress, two different directions were adopted. First, a larger dataset were considered for experiments and second, other machine learning techniques may be used for analysis.

In the study Mittal et. al. compared the outcomes of six different classifiers to identify those students who have high probability to have reappeared in some of the courses, by doing predictive analysis [19]. The investigation in this paper differs from previous related works in respect that the authors compared the accuracy of six EDM(Educational Data Mining) techniques to identify those students, who are likely to have reappear in course at an early stage, so that action can be taken to minimize the rate of failure at the end of course. Specifically, the six EDM techniques ("Naïve Bayes", "Logistic Regression", "Multilayer Perception", "Support Vector Machine", "J48 decision tree" and "Random Forest") were compared to conduct the study. The research results allows to conclude that the analysed classification algorithms are quite effective for early identification of those students who can have reappear in course, which can then be useful for parents or teachers to take a corrective action at an early stage. The prediction of student's performance can be done to a good extent by using the six features gender, caste, percentage, family income, mother's occupation and number of friends.

The two classifiers, multi-layer perception and Random Forest prove to be the most appropriate classifiers for predicting student's performance. For future work, this analysis can be further taken forward by using data sets from different universities and also applying other data pre-processing techniques.

Ravindra concluded that in the future AI will enable new capabilities to address library user's information needs.Libraries can use AI tools to provide not just information but deep intelligence offeringInsight As A Service (IAAS) [20]. It is necessary to understand the advantages and disadvantagesof artificial intelligence and machine learning for better use and its application in librariesand information centers. AI is impacting the way information is processed and searched forand information professionals will be able to use these exciting new technologies to enhancetheir services and help users find and access specific information more easily and quickly.Library and information centers will get benefits by the development of the efficient expertsystem for technical services as well as for information processing and management.

3 Research gap

Research in student performance analysis using machine learning has seen significant advancements, but there are still several gaps that offer opportunities for further investigation and innovation. Here are some potential research gaps in this area which will be covered in my research work:

3.1Fine-grained analysis of student performance: While many studies have focused on predicting overall student performance in exams, there's a need for more granular analysis. Research could explore predicting performance on specific field like the test completion course and analyzing different learning power through the differentscore like the math

score, reading and writing sorewhich will be done in my research work and it could provide valuable insights for personalized learning interventions.

3.2Incorporating diverse data sources: Most existing studies primarily use academic records and demographic information for predicting student performance. There's a gap in research exploring the integration of diverse data sources such as socio-economic factors like the impact of parental education on student's performance which will be done in my work to analyse how parent education impacts on the student's performance. Thebehavioral data on the basis of the gender can also be the factor to analyse the performance of the students in examination. This analysis will also be done in my work which could provide the prediction accuracy and gain deeper insights into student learning processes.

4 Objectives

Analysing student's performance using machine learning techniques via Python can be incredibly beneficial in educational settings. Given below is the objectives of using machine learning to analyse student's performance:

4.1Prediction:

- Forecasting students' future performance based on historical data.
- Identifying student's performance earlier to provide timely interventions.

4.2Personalized Learning:

- Tailoring educational content to individual student needs.
- Recommending resources and activities based on learning styles.

4.3Identifying Patterns:

- Discovering patterns and trends in student performance.

- Recognizing factors that influence academic success.

4.4Improving Teaching Methods:

- Enhancing teaching methodologies based on identified patterns.
- Providing insights to educators to improve their teaching strategies.

4.5Optimizing Curriculum:

- Analysing the effectiveness of different parts of the curriculum.
- Modifying or enhancing the curriculum based on identified weaknesses or strengths.

4.6Resource Allocation:

- Optimizing resource allocation based on identified needs.
- Allocating resources efficiently to support struggling students.

4.7Quality Assurance:

- Monitoring and maintaining the quality of education.
- Evaluating the effectiveness of educational programs.

5 Scope of work:

5.1Data Collection: Gathering various types of data such as gender, race/ethnicity, parental education, test preparation course, math, reading, reading and writing etc.Data can be collected from various sources like learning management systems, surveys, and historical academic records.

5.2Data Pre-processing: Cleaning and pre-processing the collected data to remove noise and inconsistencies. Handling missing data and outliers.

5.3Feature Selection/Extraction: Identifying relevant features that can influence student performance.Extracting meaningful features from the data.

5.4Model Development: Employingsupervised learning technique using linear regression.InSupervised learning technique, the training data provided to the machines work as the supervisor that teaches the machines to predict the output correctly. It applies the same concept as a student

learns in the supervision of the teacher.The aim of a supervised learning algorithm is to find a mapping function to map the input variable(x) with the output variable(y).Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between the dependent variable and one or more independent features by fitting a linear equation to observed data.

5.5 Model Evaluation: Assessing the performance of the model using appropriate metrics. The cross-validation and hyperparameter tuning to ensure the robustness of the model.

5.6 Visualisation and Interpretation: Visualizing the results with the help of powerful library module of python mainly matplotlib and seaborn for easy interpretation. With the help of Visualisation understanding the factors that influence student performance.

5.7 Integration with present educational systems to give the industrial use of the project: The scope of the proposed system "Student's Performance Analysis in Examination" is to integrate the developed models with existing educational system which will help the teachers and the educators to better analyse the student's performance and work in proper direction for the betterment in the child development. As we know in modern way of teaching and learning methodology in schools and colleges teachers and educators keep the records of all kind of performance of the students and on the basis of their performance their further training and teaching methodology works. As we know that if in a class there are 30 students and each and every student have their own level of learning capabilities and different level of potential from the very childhood.This type of implementation of performance is done in international school of India and in abroad.This work of implementation will be shown in my thesis.

5.8 Privacy and Ethics: The work will ensure the privacy and security of student data.

6 Hypothesis

The research hypotheses are based on theoretical reasoning and results from previous studies, as explained in the Literature Survey section of this study. Given these considerations, we formed the following hypotheses whose work will be proved in my thesis.

Hypothesis 1: The performance of the students is better if the parents have a better level of education. This hypothesis I will be proven in my thesis because we know that the education of the parents impact the performance of the students in examination.

Hypothesis 2: In checking the performance of test course which is divided in math, reading and writing score. We know both the gender male and female have almost same IQ level but to many other different research and observation we come to know that thatmale have better reasoning power than the female. This hypothesis will be shown in my research thesis work.

7 Research Methodology

The research methodology used in my research is the experimental approach.Analysing student performance in examinations using machine learning is an insightful approach to understand various factors that contribute to academic success. Here's an outline of an experimental approach being considered in my study:

7.1 Statement of Research Problem

In the realm of education, understanding the multifaceted determinants of student performance in examinations is pivotal for devising effective interventions to enhance learning outcomes. This study aims to leverage machine learning techniques to comprehensively analyse the diverse array of

factors influencing student performance in examinations. By delving into datasets encompassing student demographics, academic history, socio-economic background and test preparation course and other affecting factors too. This research endeavors to develop predictive models capable of discerning patterns and elucidating key predictors of academic success.

7.2Research Design

The research design which will be used in my research work to analyse the student's performance in examination will be **experimental design** in which one or more independent variables like parental level of education, race/ethnicity gender and lunch is used to observe their effect on a dependent variable like test preparation course completed or not and math, reading and writing score. It allows to establish cause-and-effect relationships between variables.

7.3Sampling design:

We will be using random sampling technique to collect student's data such as gender, race/ethnicity, parental level of education, lunch, test preparation course, math score, reading score and writing score which are the affecting factors for the analysis of student performance.

7.4Data Collection Strategy :

We have been using secondary data in form of datasets from Kaggle for machine learning which offers several benefits like diverse range of datasets ,high-quality data, large dataset repository, data exploration, visualization and educational resources. We have taken the datasets of students of 1000 records which is csv file (Excel format).

7.4.1 Tools & Techniques

In research, tools and techniques serve various purposes, aiding in data collection, analysis, visualization, and interpretation. Analyzing student

performance in exams involves a variety of tools and techniques. In my research the tool used are:

IDE(Integrated Development Environment For Programming)–

Anaconda3 Navigator offers several benefits for machine learning (ML) practitioners specially for package management, pre-installed libraries, cross-platform compatibility, virtual environments and community support. Anaconda includes popular IDEs like Jupyter Notebook, which provide powerful tools for developing, debugging, and visualizing machine learning models.

7.4.2 Platform:Jupyter Notebook for Python Programming (Version 6.3.0):

Jupyter Notebook is a widely used interactive computing environment that allows users to create and share documents containing live code, equations, visualizations, and narrative text. Jupyter Notebook, particularly version 6.3.0, is beneficial for machine learning programming.Some of the interactive features of Jupyter Notebook are interactive development, visualization, documentation, code reusability, experimentation & prototyping, integration with data analysing tools.

7.4.3 Computer programming in Python Version 3:

Python 3 is particularly useful for machine learning programming due to its rich ecosystem, ease of use, community support, flexibility and easy integration as python seamlessly integrates with other technologies commonly used in machine learning workflows, such as data manipulation libraries.Overall, Python's combination of simplicity, versatility, and powerful libraries makes it an ideal choice for machine learning programming.

7.4.4 Importing libraries & datasets

In [2]:

```
#import basic libraries and load the dataset
```

In [2]:

```
# for visualizations
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
sns.set_style("whitegrid")
plt.style.use("fivethirtyeight")
```

In [3]:

```
data = pd.read_csv('StudentsPerformance.csv')
```

In [4]:

```
data.head()
```

Out[4]:

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	female	group B	bachelor's degree	standard	none	72	72	74
1	female	group C	some college	standard	completed	69	90	88
2	female	group B	master's degree	standard	none	90	95	93
3	male	group A	associate's degree	free/reduced	none	47	57	44
4	male	group C	some college	standard	none	76	78	75

Implementation of different Python libraries:

Pandas libraries: To work with data in the form of table, data frame we need to import pandas library that provides high-performance data manipulation in Python. Pandas Data Frame is a widely used data structure which works with a two-dimensional array with labeled axes (rows and columns).

Numpy libraries: Numpy provides a convenient and efficient way to handle the vast amount of data. Numpy is also very convenient with Matrix multiplication and data reshaping. Numpy is fast which makes it reasonable to work with a large set of data.

Matplotlib:Matplotlib is a Python library which is defined as a multi-platform data visualization library built on Numpy array.

Seaborn : Seaborn is one of library for visualization of the graphical statistical plotting in Python. Seaborn provides many color palettes and defaults styles to make the creation of many statistical plots in Python.

7.4.5 Data planning & data analysis

Data planning and data analysis are critical components of any research project, ensuring that the data collected is meaningful and can be effectively analyzed to draw valid conclusions. Here's a breakdown of each:

7.4.5.1 Data Preprocessing :

Data pre-processing include cleaning of the data and prepare it for analysis. This process involves handling missing values, outliers, and encoding categorical variables if necessary.

7.4.5.2 Feature Selection/Engineering:

Identify which features are most relevant to predicting student performance. We might use techniques like correlation analysis or feature importance from machine learning models.

7.4.5.3 Model Selection:

Choose appropriate machine learning models for the analysis is an important part. Models for this type of analysis include in my research is:

Linear Regression: For predicting continuous outcomes like examination scores.

Classification models: For predicting categorical outcomes like pass/fail or grade categories.

7.4.5.4 Training and Testing:

Split the data into training and testing sets. Train the chosen models on the training set and evaluate their performance on the testing set using appropriate metrics (e.g., accuracy, precision, recall, etc.).It is called Train/Test because you split the data set into two sets: a training set and a testing set. Mainly we split the data in the ratio of 80% for training, and 20% for testing. The steps included in my research for training and testing data will be:

We will train the model means creating the model for the training set.

Finally we will test the model means testing the accuracy the testing set.

The above two steps will be repeated until we get more than 95% of accuracy.

7.5.5.5 Model Evaluation and Interpretation:

Evaluate the performance of the model and interpret the results. Identify which factors are most influential in predicting student performance. We can use techniques like feature importance plots or values for model interpretability.

8 Conclusion

To achieve our goal my thesis will have sections in form of chapters namely the **Introduction,Literature Review, Conceptual Framework, Plan & Procedure of the Study, The Analysis of Data, Finding and the Interpretation of Data and Conclusion of Studies.**

Finally, performance analysis for students are a major issue in educational system. It is important that they are countered. The work reported in this synopsis indicates the machine learning technique in which supervised learning algorithm with respect to student records will be used as supervised learning is more convenient way to analyse data with respect to train and test data. With recent advancements in data acquisition systems and system performance indicates, educational systems are now studied more effectively yet with much less effort. State-of-the-art data mining and machine learning techniques will be proposed for analysing and monitoring massive data giving rise to a whole new field of big data analytics. However, only a fewstudies proposed remedial solutions to provide in-time feedback to students, instructors,and educators to address the problems. In later part of my work will focus more on developingan efficient ensemble method to practically deploy the ML-based performance predictionmethodology and search for dynamic ways or methods to predict students' performanceand provide automatic needed remedial actions to help the students as early as possible. More emphasize in my research will be given on using appropriate machine learning techniques in predicting student's performance. We willimplement

some of the excellent works and focus more on dynamic nature of student's performance.

9. References

1. Kasi, U., &Prasad, S. (2020). Hypothesis-Testing Factors Affecting Students' Academic Performance. India.*International Journal of Computer Applications (0975 – 8887*), *Volume 175* – No. 30.
2. Durge, D.,Bagul, N., &Bhavsar, S.(2020). STUDENT PERFORMANCE ANALYSIS SYSTEM. *JETIR, Volume 7*, Issue 9 ,*www.jetir.org (ISSN-2349-5162)*
3. Reddy, H. K.,Harinath., Lokesh., Suresh.,Md. Jawad.,&Reddy, K.V.(2021).Student Performance Analysis System using Machine Learning.*Journal of Resource Management and Technology.ISSN NO: 0745-6999.Vol* 12, Issue3.
4. Karthikeyan,R.,Babu,S.S., &Gokulakrishnan,P. (2021). Machine Learning based Student Performance Analysis System. *IT in Industry, vol.* 9, *no.1. Published online 17-March-2021.SSN (Print): 2204-0595.ISSN (Online): 2203-1731*
5. Jirage, K.,Choudhari,R., Gupta,V.,Patthe,A. &Maind,D.(2015)Student Performance Analysis Based on Machine Learning." *INTERNATIONAL JOURNAL FOR RESEARCH IN EMERGING SCIENCE AND TECHNOLOGY, VOLUME*-2, *ISSUE-3,E-ISSN: 2349-7610.*
6. Prajapati, A.P., Sharma,S.K., &Sharma,M.K.(2017).Student's Performance Analysis Using Machine Learning Tools.*International Journal of Scientific & Engineering Research* Volume 8, Issue 10,*ISSN 2229-5518.*
7. Anusha,M., Kartik.,Padmini,P.Rani., &Vsrikanth.(2019). Prediction of Student Performance using Machine Learning.*International Journal of*

Engineering and Advanced Technology (IJEAT).ISSN: 2249 – 8958, Volume-8 *Issue-6, August 2019.Retrieval Number E7520068519 /2019©BEIESP*. DOI: 10.35940/ijeat.E7520.088619.

8. Kumar, M., &Singh, A.J. (2019).Performance Analysis of Students Using Machine Learning & Data Mining Approach.*International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume*-8 Issue-3.
9. Naveen,L., &Chandrakanth,P.(2022).Student's Performance Analysis Using Machine Learning Tools. *Journal of Engineering Sciences Vol* 13 *Issue 08,2022,ISSN:0377-9254.*
10. Pallathadka, H.,Wenda.,Ramirez-Asís,A.E., Asís-López,M., Flores-Albornoz, J., &Phasinam,K.(2021). Classification and prediction of student performance data using various machine learning algorithms.*DOI: 10.1016/j.matpr.2021.07.382.Journal homepage: www.elsevier.com/locate/matpr.*
11. Acharya,A.(2017).Early Prediction of Students Performance using MachineLearningTechniques.*International Journal of Computer Applications (0975 – 8887) Volume* 107 – No.1.
12. Priya,S., Ankit,T., &Divyansh.(2021).Student Performance Prediction Using Machine Learning.*Advances in Parallel Computing Technologies and Applications D.J. Hemanth et al. (Eds.) © 2021 The authors and IOS Press. doi:10.3233/APC210137.*
13. Dhilipan,J., Vijayalakshmi,N., Suriya,S., &Christopher,A. (2020).Prediction of Students Performance using Machine Learning.*IVC RAISE 2020 ,IOP Conf. Series: Materials Science and Engineering 1055. 012122, IOP Publishing, doi:10.1088/1757-899X/1055/1/012122* .
14. Alshabandar,R., Hussain,A.,Keight, R., &Khan,W. (2020) Students Performance Prediction in Online Courses Using Machine Learning Algorithms. *IVC RAISE 2020 ,IOP Conf. Series: Materials Science and*

Engineering 1056. 012123, IOP Publishing, doi:10.1088/1757-899X/1055/1/012123 .

15. Kotaiah, B., &FazalulHaque,S.M.(2019). Analysis of Student Result using Machine Learning in Python. *Turkish Journal of Computer and Mathematics Education* Vol.10 No. 01*, 467-473.*
16. Shahiri, A.M., Husain, W., &Abdul Rashid,N. (2017). A Review on Predicting Student's Performance using Data Mining Techniques.*Procedia Computer Science 72.414 – 422, Elsevier.*
17. Memon,M.M., Jameel, S.M., Hashmani,M.A., Junejo,S., &Raza,K.(2020). Analysis of Student Procrastinatory Behavior in Virtual Learning Environments Using Machine Learning. *Journal of Hunan University (Natural Sciences),Vol.* 47. No. 10.
18. Gupta,Y., Raghuwanshi,G., &Gopalakrishna.(2019).Student Performance Evaluation Expert System using Machine learning to make Cognitive decisions. *SUSCOM. Hosting by Elsevier SSRN.*
19. Mittal, S., Aggarwal,D., &Bali,V. (2019).Prediction Model for Classifying Students Based on Performance using Machine Learning Techniques. *International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume*-8, *Issue-2S7.*
20. Ravindra,D.,Sarode.K.,Chhaya,A., &Khanzode. (2020).Advantage and disadvantages of Artificial Intelligence and Machine Learning :A Literature Review.*International Journal of Library & Information Science (IJLIS).Volume* 9, *Issue 1, pp. 30-36, Article ID: IJLIS_09_01_004.*